2X

**УДК 519.2**

***Ю.С. Харин, А.И. Петлицкий, М.В. Мальцев***
НИИ прикладных проблем математики и информатики БГУ, г. Минск, Беларусь
kharin@bsu.by, piatlitski@bsu.by, maltsew@mail.ru

# Выявление зависимостей большой глубины на основе марковских моделей

Построены два статистических теста для выявления зависимости в случайной последовательности и обнаружения отклонений вероятностного распределения элементов последовательности от равномерного. Первый тест основан на частотных статистиках цепи Маркова *s*-го порядка с *r* частичными связями, второй – на частотных статистиках цепи Маркова переменной длины. Представлены результаты компьютерных экспериментов.

## Введение

Выявление зависимости в случайной последовательности и обнаружение отклонений вероятностного распределения элементов последовательности от равномерного являются важнейшими проблемами в защите информации [1-3], генетике [4] и других приложениях. Обзор существующих методов решения этих задач представлен в [2]. Актуальность проблемы построения новых статистических тестов [5] связана с тем, что: 1) многие известные тесты проверяют лишь одно из вероятностных свойств, характеризующих случайную последовательность; 2) большинство тестов построено «эвристически» и не фиксирует семейство альтернатив; 3) многие из существующих тестов не имеют теоретических оценок мощности.

В данной статье разработаны два новых теста для статистической проверки гипотезы $H_0 =${наблюдаемая последовательность есть равномерно распределенная случайная последовательность (РРСП)} против альтернативы $H_1 = \overline{H}_0$; РРСП – это случайная последовательность, элементы которой независимы в совокупности и имеют равномерное распределение вероятностей [2]. Первый тест $T_{ЦМ(s,r)}$ основан на частотных статистиках новой марковской модели – цепи Маркова *s*-го порядка с *r* частичными связями ЦМ(*s*,*r*) [6], а второй тест $T_{ЦМПД}$ – на частотных статистиках цепи Маркова переменной длины [7]. Для тестов $T_{ЦМ(s,r)}$ и $T_{ЦМПД}$ исследована мощность для семейства контигуальных альтернатив, а также проведено сравнение с тестом $T_{ЦМ}$ на основе частот цепи Маркова *s*-го порядка [8].

## Тест, основанный на частотных статистиках цепи Маркова с частичными связями

Обозначим: $A = \{0,1,\dots,N-1\}$ – множество состояний мощности $2 \le N < \infty$; $J_i^k = (j_i, j_{i+1},\dots, j_k) \in A^{k-i+1}$ – мультииндекс $(k-i+1)$-го порядка, $k \ge i$; $\{x_t \in A\}$ – однородная стационарная цепь Маркова *s*-го порядка с вероятностями одношаговых переходов

$$p_{j_1,\dots,j_s,j_{s+1}} = \mathrm{P}\{x_{t+s} = j_{s+1} \mid x_{t+s-1} = j_s,\dots,x_t = j_1\},\ J_1^{s+1} \in A^{s+1},\ t \ge 1;$$

$r \in \{1,2,...,s\}$ – параметр, называемый числом связей; $M_r^0 = \left(m_1^0, m_2^0, ..., m_r^0\right)$ – целочисленный $r$-вектор с упорядоченными компонентами $1 = m_1^0 < m_2^0 < \ldots < m_r^0 \le s$, называемый шаблоном связей; $Q = \left(q_{J_1^{r+1}}\right)_{J_1^{r+1} \in A^{r+1}}$ – некоторая ($r$+1)-мерная стохастическая матрица.

Цепь Маркова $\{x_t\}$ называется цепью Маркова $s$-го порядка с $r$ частичными связями и обозначается ЦМ($s$,$r$) [6], если ее вероятности одношаговых переходов имеют вид:

$$p_{j_1,\ldots,j_s,j_{s+1}} = q_{j_{m_1^0},\ldots,j_{m_r^0},j_{s+1}}, \; J_1^{s+1} \in A^{s+1}. \tag{1}$$

Соотношение (1) означает, что вероятность перехода процесса $x_t$ в состояние $j_{s+1}$ зависит не от всех $s$ предыдущих состояний процесса $j_1,...,j_s$, а лишь от $r$ избранных состояний $j_{m_1^0},...,j_{m_r^0}$. Если $r = s$, то получаем цепь Маркова $s$-го порядка [9].

Примем еще несколько обозначений: $X_1^n = (x_1, x_2, \ldots, x_n)$ – наблюдаемая реализация длительности $n$; $\delta_{i,k}$ – символ Кронекера;

$$\nu_{ЦМ(s,r)}\left(J_1^{r+1}; M_r\right) = \sum_{t=1}^{n-s}\left(\prod_{i=1}^{r}\delta_{x_{t+m_i-1},j_i}\right)\delta_{x_{t+s},j_{r+1}}, \; J_1^{r+1} \in A^{r+1}, \tag{2}$$

– частотные статистики цепи Маркова ЦМ($s$,$r$);

$$\mu_{ЦМ(s,r)}\left(J_1^{r+1}; M_r^0\right) = \mathrm{P}\left\{x_{t+m_1^0-1} = j_1, ..., x_{t+m_r^0-1} = j_r, x_{t+s} = j_{r+1}\right\}, \; J_1^{r+1} \in A^{r+1},$$

– распределение вероятностей $(r+1)$-грамм цепи Маркова с частичными связями; $\hat{\mu}_{ЦМ(s,r)}\left(J_1^{r+1}; M_r^0\right) = \nu_{ЦМ(s,r)}\left(J_1^{r+1}; M_r^0\right)/(n-s)$ – несмещенная и состоятельная частотная оценка вероятности $\mu_{ЦМ(s,r)}\left(J_1^{r+1}; M_r^0\right)$, $J_1^{r+1} \in A^{r+1}$.

Построим тест проверки гипотез $H_0$: $\{x_t\}$ – РРСП, то есть $q_{J_1^{r+1}} = N^{-1}$, $J_1^{r+1} \in A^{r+1}$; $H_{1,ЦМ(s,r)}$: $\{x_t\}$ – цепь Маркова ЦМ($s,r$), для которой матрица $Q$ имеет вид:

$$q_{J_1^{r+1}} = q_{J_1^{r+1}}(n) = \frac{1}{N}\left(1 + b_{J_1^{r+1}} / \sqrt{n-s}\right) > 0, \; J_1^{r+1} \in A^{r+1}, \tag{3}$$

где $\sum_{j_{r+1} \in A} b_{J_1^{r+1}} = 0$, $\sum_{J_1^{r+1} \in A^{r+1}} | b_{J_1^{r+1}} | \ne 0$. Соотношение (3) определяет контигуальное семейство альтернатив [10] и означает, что при увеличении длительности $n$ наблюдаемой последовательности, гипотеза $H_1$ сближается с $H_0$ со скоростью $O\left(n^{-1/2}\right)$.

Обозначим

$$\xi_{ЦМ(s,r)}\left(J_1^{r+1}\right) = \sqrt{(n-s)N^{r+1}}\left(\hat{\mu}_{ЦМ(s,r)}\left(J_1^{r+1}; M_r^0\right) - N^{-(r+1)}\right), \; J_1^{r+1} \in A^{r+1}, \tag{4}$$

$$\rho_{ЦМ(s,r)} = \sum_{J_1^r \in A^r}\left(\sum_{j_{r+1} \in A}\xi^2_{ЦМ(s,r)}\left(J_1^{r+1}\right) - \frac{1}{N}\left(\sum_{j_{r+1} \in A}\xi_{ЦМ(s,r)}\left(J_1^{r+1}\right)\right)^2\right). \tag{5}$$

**Теорема 1.** При $n \to \infty$ случайная величина $\rho_{ЦМ(s,r)}$ в случае справедливости гипотезы $H_0$ имеет $\chi^2$-распределение с $U = N^r(N-1)$ степенями свободы.

При помощи теоремы 1 строится тест $T_{ЦМ(s,r)}$ [11] для проверки гипотез $H_0$ и $H_{1,ЦМ(s,r)}$, основанный на частотных статистиках цепи Маркова с $r$ частичными связями:

1) по наблюдаемой последовательности $X_1^n$ длительности $n$ строятся частотные статистики $\left\{\nu_{ЦМ(s,r)}\left(J_1^{r+1}; M_r\right): J_1^{r+1} \in A^{r+1}\right\}$ согласно (2);

2) согласно (4), (5) вычисляются статистики $\left\{\xi_{ЦМ(s,r)}\left(J_1^{r+1}\right): J_1^{r+1} \in A^{r+1}\right\}$ и $\rho_{ЦМ(s,r)}$;

3) вычисляется P-значение: $P = 1 - G_U\left(\rho_{ЦМ(s,r)}\right)$, где $G_U(\cdot)$ – функция $\chi^2$-распределения с $U$ степенями свободы;

4) решение выносится с помощью решающего правила:

принимается $\{H_0$, если $P > \varepsilon$; $H_{1,ЦМ(s,r)}$, если $P \le \varepsilon\}$,

где $\varepsilon \in (0,1)$ – заданный уровень значимости теста.

**Теорема 2.** При $n \to \infty$ случайная величина $\rho_{ЦМ(s,r)}$ в случае справедливости гипотезы $H_{1,ЦМ(s,r)}$ имеет нецентральное $\chi^2$-распределение с $U$ степенями свободы и параметром нецентральности

$$a_{ЦМ(s,r)} = \frac{1}{N^{r+1}} \sum_{J_1^{r+1} \in A^{r+1}} b_{J_1^{r+1}}^2 . \qquad (6)$$

***Следствие 1.*** Мощность теста $T_{ЦМ(s,r)}$ при $n \to \infty$ удовлетворяет асимптотическому соотношению:

$$w \to 1 - G_{U,a_{ЦМ(s,r)}}\left(G_U^{-1}(1-\varepsilon)\right),$$

где $G_{U,a_{ЦМ(s,r)}}(\cdot)$ – функция нецентрального $\chi^2$-распределения с U степенями свободы и параметром нецентральности $a_{ЦМ(s,r)}$, определяемым (6).

***Следствие 2.*** Тест $T_{ЦМ(s,r)}$ имеет большую мощность по сравнению с тестом $T_{ЦМ}$.

# Тест, основанный на частотных статистиках цепи Маркова переменной длины

Цепь Маркова $\{x_t\}$ называется цепью Маркова переменной длины порядка $s$ [7], если ее вероятности одношаговых переходов имеют вид:

$$p_{j_1,\dots,j_s,j_{s+1}} = q_{j_{s-l+1},\dots,j_s,j_{s+1}}, \quad l = l\left(J_1^s\right), \quad J_1^{s+1} \in A^{s+1}. \qquad (7)$$

Соотношение (7) означает, что вероятность перехода в состояние $j_{s+1}$ зависит не от всех $s$ предыдущих состояний, а лишь от $l = l\left(J_1^s\right)$ предыдущих состояний. Если $l\left(J_1^s\right) = s$, то получаем цепь Маркова $s$-го порядка [9].

Функция $l(\cdot)$ определяется с помощью контекстной функции $c\left(J_1^s\right) = J_{s-l+1}^s$, $l\left(J_1^s\right) = \left|c\left(J_1^s\right)\right|$, $J_1^s \in A^s$. Контекстную функцию удобно представлять в виде корневого дерева $\tau$, которое называется контекстным деревом. У каждой вершины в таком дереве может быть не более $N$ потомков, поскольку каждому узлу (кроме корня) соответствует элемент из множества состояний $A$. Каждому значению контекстной функции соответствует ветвь данного дерева.

Примем обозначения:

$$\nu_{ЦМПД}\left(J_1^{l+1}\right) = \sum_{t=1}^{n-l}\left(\prod_{i=1}^{l+1} \delta_{x_{t+i-1},j_i}\right), \; J_1^l \in \tau, \; j_{l+1} \in A, \qquad (8)$$

– частотные статистики цепи Маркова переменной длины;

$$\mu_{ЦМПД}\left(J_1^{l+1}\right)=\mathrm{P}\{x_t=j_1,\ldots,x_{t+l-1}=j_l,x_{t+l}=j_{l+1}\},\ J_1^l\in\tau,\ j_{l+1}\in A,$$

– распределение вероятностей $(l+1)$-грамм цепи Маркова переменной длины; $\hat{\mu}_{ЦМПД}\left(J_1^{l+1}\right)=\nu_{ЦМПД}\left(J_1^{l+1}\right)/(n-l)$ – несмещенная и состоятельная частотная оценка вероятности $\mu_{ЦМПД}\left(J_1^{l+1}\right)$, $J_1^l\in\tau$, $j_{l+1}\in A$.

Построим тест проверки гипотез $H_0$: $\{x_t\}$ – РРСП, то есть $q_{J_1^{l+1}}=N^{-1}$, $J_1^l\in\tau$, $j_{l+1}\in A$; $H_{1,ЦМПД}$: $\{x_t\}$ – цепь Маркова ЦМ($s$,$r$), для которой матрица $Q$ имеет вид аналогичный (3):

$$q_{J_1^{l+1}}=q_{J_1^{l+1}}(n)=\frac{1}{N}\left(1+b_{J_1^{l+1}}/\sqrt{n-s}\right)>0,\ J_1^l\in\tau,\ j_{l+1}\in A, \tag{9}$$

где $\sum_{j_{l+1}\in A}b_{J_1^{l+1}}=0$, $\sum_{J_1^l\in\tau,\,j_{l+1}\in A}|b_{J_1^{l+1}}|\neq 0$. Соотношение (9) определяет контигуальное семейство альтернатив [10].

Определим случайные величины:

$$\xi_{ЦМПД}\left(J_1^{l+1}\right)=\sqrt{(n-s)N^{l+1}}\left(\hat{\mu}_{ЦМПД}\left(J_1^{l+1}\right)-N^{-(l+1)}\right),\ J_1^l\in\tau,\ j_{l+1}\in A, \tag{10}$$

$$\rho_{ЦМПД}=\sum_{J_1^l\in\tau}\left(\sum_{j_{l+1}\in A}\xi^2_{ЦМПД}\left(J_1^{l+1}\right)-\frac{1}{N}\left(\sum_{j_{l+1}\in A}\xi_{ЦМПД}\left(J_1^{l+1}\right)\right)^2\right) \tag{11}$$

**Теорема 3.** При $n\to\infty$ случайная величина $\rho_{ЦМПД}$ в случае справедливости гипотезы $H_0$ имеет $\chi^2$-распределение с $U=|\tau|\left(N-1\right)$ степенями свободы.

При помощи теоремы 3 строится тест $T_{ЦМПД}$ для проверки гипотез $H_0$ и $H_{1,ЦМПД}$, основанный на частотных статистиках цепи Маркова с $r$ частичными связями:

1) по наблюдаемой последовательности $X_1^n$ длительности $n$ строятся частотные статистики $\left\{\nu_{ЦМПД}\left(J_1^{l+1}\right):J_1^l\in\tau,\ j_{l+1}\in A\right\}$ согласно (8);

2)согласно (10), (11) вычисляются статистики $\left\{\xi_{ЦМПД}\left(J_1^{l+1}\right):J_1^l\in\tau,\ j_{l+1}\in A\right\}$ и $\rho_{ЦМПД}$;

3) вычисляется Р-значение: $P=1-G_U\left(\rho_{ЦМПД}\right)$, где $G_U(\cdot)$ – функция $\chi^2$-распределения с $U$ степенями свободы;

4) решение выносится с помощью решающего правила:

принимается $\{H_0$, если $P>\varepsilon$; $H_{1,ЦМПД}$, если $P\le\varepsilon\}$,

где $\varepsilon\in(0,1)$ – заданный уровень значимости теста.

**Теорема 4.** При $n\to\infty$ случайная величина $\rho_{ЦМПД}$ в случае справедливости гипотезы $H_{1,ЦМПД}$ имеет нецентральное $\chi^2$-распределение с $U$ степенями свободы и параметром нецентральности

$$a_{ЦМПД}=\frac{1}{|\tau|N}\sum_{J_1^l\in\tau,\,j_{l+1}A}b^2_{J_1^{l+1}}. \tag{12}$$

***Следствие 3.*** Мощность теста $T_{ЦМПД}$ при $n\to\infty$ удовлетворяет асимптотическому соотношению:

$$w\to 1-G_{U,a_{ЦМПД}}\left(G_U^{-1}(1-\varepsilon)\right),$$

где $G_{U,a_{ЦМПД}}(\cdot)$ – функция нецентрального, $\chi^2$-распределения с U степенями свободы и параметром нецентральности $a_{ЦМПД}$ определяемым (12).

***Следствие 4.*** Тест $T_{ЦМПД}$ имеет большую мощность по сравнению с тестом $T_{ЦМ}$.

# Численные результаты

Проведены численные эксперименты на модельных и реальных данных.

***Пример 1 (модельные данные).*** На рис. 1 представлена зависимость мощности тестов $T_{ЦМ(s,r)}$, $T_{ЦМ}$ для альтернативы $H_{1,ЦМ(s,r)}$ от $n$ при $\varepsilon = 0,05$, $N = 4$, $s = 6$, $r = 4$, $M_r^0 = (1,4,5,6)$ и матрице $Q$, для которой: 1) $b_{J_1^r,0},\dots,b_{J_1^r,N-2}$ генерировались с помощью стандартного генератора равномерно распределенных на [-13,13] псевдослучайных чисел, а $b_{J_1^r,N-1} = -(b_{J_1^r,0} + \dots + b_{J_1^r,N-2})$; 2) функция нецентрального $\chi^2$-распределения имеет $U = 768$ степеней свободы и параметр нецентральности $a_{ЦМ(s,r)} = 138,5$. На этом рисунке квадратиками и кружками указаны значения оценки мощности $\hat{w}$ для $T_{ЦМ(s,r)}$ и $T_{ЦМ}$ соответственно, полученные с помощью метода Монте-Карло при числе прогонов, равном 1000; пунктирные линии – верхняя и нижняя 99 % доверительные границы для мощности; сплошная линия – теоретическое значение $w$, найденное в следствии 1. Из рис. 1 видно, что для указанных значений параметров мощность теста $T_{ЦМ(s,r)}$ приблизительно в 4 раза превосходит мощность теста $T_{ЦМ}$, что согласуется со следствием 2. Отметим, что при $n \to \infty$ мощность тестов не стремится к 1, так как при увеличении $n$ гипотеза $H_1$ сближается с $H_0$ (контигуальная постановка задачи).

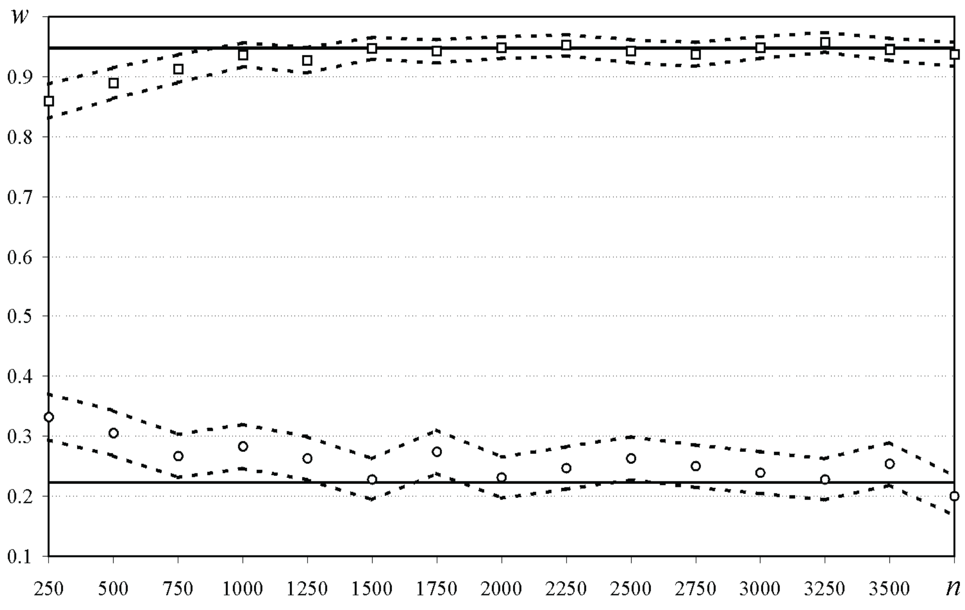



Рисунок 1 – Зависимость мощности от $n$

***Пример 2 (модельные данные).*** Исследовалась зависимость мощности тестов $T_{ЦМПД}$, $T_{ЦМ}$ для альтернативы $H_{1,ЦМПД}$ от $n$ при $\varepsilon = 0,05$, $N = 2$, $s = 8$ и контекстном дереве $\tau$, представленном на рис. 2. Эта зависимость проиллюстрирована на рис. 3 (квадратики и кружки – значения оценки мощности $\hat{w}$ для $T_{ЦМПД}$ и $T_{ЦМ}$ соответственно; пунктирные линии – верхняя и нижняя 99 % доверительные границы для мощности; сплошная линия – теоретическое значение $w$, определяемое следствием 3). Из рис. 3 видно, что для этих значений параметров мощность теста $T_{ЦМПД}$ приблизительно в 3 раза превосходит мощность теста $T_{ЦМ}$, что согласуется со следствием 4.

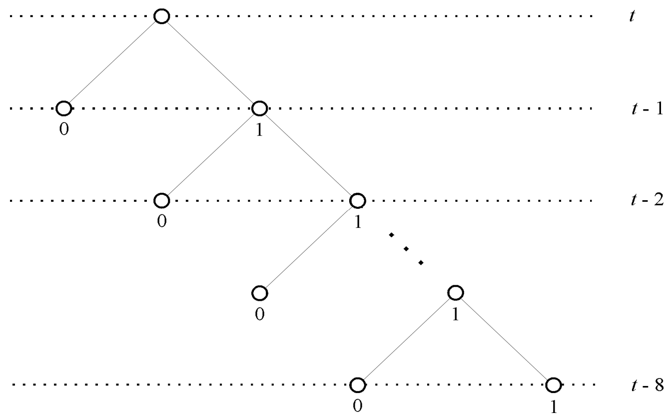



Рисунок 2 – Контекстное дерево $\tau$

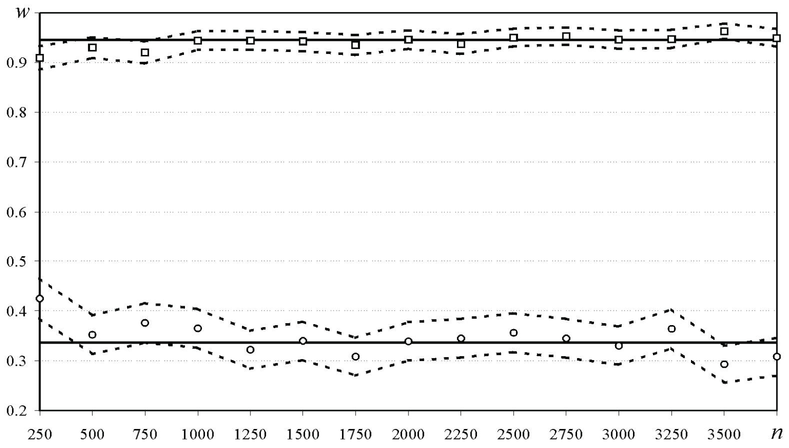



Рисунок 3 – Зависимость мощности от $n$

***Пример 3 (реальные данные).*** Исследовался генератор псевдослучайных последовательностей А5/1 [12], состоящий из трех коротких линейных регистров сдвига с обратной связью. Алгоритм А5/1 используется в сети GSM для обеспечения защиты информации на уровне базовая-мобильная станция.

При тестировании генератора А5/1 с помощью теста $T_{ЦМ(s,r)}$ его выходная последовательность разбивалась на 12-битовые фрагменты, и каждый такой фрагмент рассматривался как буква алфавита $A = \{0,1,...,2^{12} - 1\}$, мощности $N = 2^{12}$. На вход теста $T_{ЦМ(s,r)}$ поступали 250 реализаций выходной последовательности длительности $n$ бит каждая, сгенерированных этим криптоалгоритмом; параметры теста: $s = 2$, $r = 1$, $M_r^0 = (1)$, $\varepsilon = 0,05$. Результаты исследований приведены в табл. 1. Для РРСП при уровне значимости $\varepsilon = 0,05$ среднее число отклоненных из 250 реализаций равнялось бы 12,5. Таким образом, представленные в табл. 1 результаты свидетельствуют о сильной неслучайности выходной последовательности алгоритма А5/1.

Таблица 1 – Результаты тестирования генератора А5/1

| Длина последовательности $n$, бит | $4 \cdot 2^{21}$ | $5 \cdot 2^{21}$ | $6 \cdot 2^{21}$ | $7 \cdot 2^{21}$ | $8 \cdot 2^{21}$ |
|---|---|---|---|---|---|
| Количество (частота) отклоненных тестом $T_{ЦМ(s,r)}$ реализаций | 37 (0,148) | 59 (0,236) | 72 (0,288) | 91 (0,364) | 105 (0,420) |

# Заключение

Построены статистические тесты на основе марковских частотных статистик, которые позволяют выявлять зависимости высокого порядка и специфической структуры. Проведенные компьютерные эксперименты на модельных и реальных данных иллюстрируют работоспособность построенных тестов.

# Литература

1. Кнут Д. Искусство программирования: В 3 т. – М.: Мир, 1992.
2. Харин Ю.С. и др. Математические и компьютерные основы криптологии. – Мн.: Новое знание, 2003.
3. Иванов М.А., Чигунков И.В. Теория, применение и оценка качества генераторов псевдослучайных последовательностей. – М.: КУДИЦ-ОБРАЗ, 2003.
4. Уотермен М.С. Математические методы анализа последовательностей ДНК. – М.: Мир, 1999.
5. Харин Ю.С., Ярмола А.Н., Петлицкий А.И. Методы и алгоритмы статистического тестирования генераторов случайных и псевдослучайных последовательностей в системах информационной безопасности // Искусственный интеллект. – 2006. – № 3. – С. 793-803.
6. Харин Ю.С. Цепи Маркова с *r*-частичными связями и их статистическое оценивание // Доклады НАН Беларуси. – 2004. – Т. 48, № 1. – С. 40-44.
7. Buhlmann P., Wyner A. Variable length Markov chains // The Annals of Statistics. – 1999. – Vol. 27, № 2. – P. 480-513.
8. Тихомирова М.И., Чистяков В.П. О двух статистиках типа хи-квадрат, построенных по частотам цепочек состояний сложной цепи Маркова // Дискретная математика. – 2003. – Т. 15, № 2. – С. 149-159.
9. Дуб Дж. Вероятностные процессы. – М.: ИЛ, 1956.
10. Руссас Дж. Контигуальность вероятностных мер. – М.: Мир, 1975.
11. Петлицкий А.И., Харин Ю.С. Проверка гипотез о независимости и равномерном вероятностном распределении элементов случайной последовательности // Вестник БГУ. – Сер. 1, 2007. – № 3. – С. 74-80.
12. Асосков А.В. и др. Поточные шифры. – М.: КУДИЦ-ОБРАЗ, 2003.

***Ю.С. Харін, А.І. Петлицький, М.В. Мальцев***

**Виявлення залежностей більшої глибини на основі марковських моделей**

Побудовані два статичні тести у випадковій послідовності і знайдення відмінностей імовірного розподілу елементів послідовності від рівномірного. Перший тест заснований на частотних статистиках мережі Маркова s-го порядку з r частковими зв’язками, другий – на частотних статистиках мережі Маркова змінної длини. Наявні результати комп’ютерних експериментів.

***Yu.S. Kharin, A.I. Piatlitski, M.V. Maltsew***

**Detection of High-Order Dependences Based on Markovian Models**

Statistical decision rules for detection of high-order dependencies and for testing of *s* dimensional uniformity of discrete time series are constructed. The first test is based on frequency statistics of Markov chain with partial connections. The second test is based on frequency statistics of variable length Markov chain. Asymptotic properties of proposed tests are found. Numerical results are given.

*Статья поступила в редакцию 02.07.2008.*